ABSTRACT OF THE DISCLOSURE

A method of generalizing a dataset having a plurality of sequences defined over a lexicon of tokens is provided. The method comprises: searching over the dataset for similarity sets, where each similarity set comprises a plurality of segments of size L having L-S common tokens and S uncommon tokens; and defining a plurality of equivalence classes corresponding to uncommon tokens of at least one similarity set. The method may further comprise a step in which a plurality of significant patterns are extracted, where each significant pattern corresponds to a most significant partial overlap between one sequence of the dataset and other sequences of the dataset. In one embodiment, a generalized dataset represented by a graph or a forest is constructed, and can be realized as a context-free grammar. The graph or forest can be used for generating sequences and/or testing grammatical structures.